ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095417

LIGHTWEIGHT FISHER VECTOR TRANSFER LEARNING FOR VIDEO DEDUPLICATION

†*Chris Henry,* †*Rijun Liao,* ⋆*Ruiyuan Lin,* ⋆*Zhebin Zhang,* ⋆*Hongyu Sun, and* †*Zhu Li*

†University of Missouri-Kansas City
⋆InnoPeak Technology (Oppo US Research Center)

ABSTRACT

Video deduplication in cloud and on devices is a key challenge for storage and communication efficiency. The lifetime of video content creation, communication/sharing, and consumption can generate multiple versions of the same content with variations in coding and editing effects. In this work, we develop a lightweight and robust deduplication feature based on the fisher vector aggregation of Scale-Invariant Feature Transform (SIFT) keypoints. The fisher vector representation is used for a deduplication transfer learning process that utilizes a lightweight Multilayer Perceptron (MLP) network with center loss to learn a compact and distinctive feature. Simulation on the CC_WEB_VIDEO dataset demonstrated that the proposed feature is extremely robust in deduplication with respect to typical editing effects and coding/transcoding degenerations while being computationally very lightweight compared to other solutions.

Index Terms— Video deduplication, near-duplicate video detection, near-duplicate video copy detection, fisher vector aggregation.

1. INTRODUCTION

In recent years, the amount of videos recorded and shared online has skyrocketed. This is primarily due to the increasing popularity of social networks and mobile devices. Consequently, the number of illegal pirate videos have also increased. Such illegal pirate videos contain the same content as the original videos along with a few subtle differences to dodge copy detection systems. These differences are generally created by adding variations like flipping, changing aspect ratio, color, frame rate, padding, overlaying text, etc. The task of detecting these near-duplicates is referred to as near-duplicate video retrieval (NDVR). Moreover, storing this enormous amount of data is a challenging issue. The knowledge about duplicate or near-duplicate videos is crucial to minimizing the amount of storage and processing needed.

Most NVDR techniques consist of feature extraction followed by computing a similarity score. The ideal goal is to generate a feature vector that is highly distinctive and lightweight. The study in [1] presented a video copy detection system that matched individual frames and verified their spatio-temporal consistency. Local patches from frames were extracted via the Hessian-Affine region detector [2] and described via Scale-Invariant Feature Transform (SIFT) [3] or CS-LBP [4] descriptors. The study in [5] represented frames using a SIFT and bag-of-words representation and used weak geometric consistency to exclude incorrect matches. Temporal-concentration SIFT (TCSIFT) was proposed in [6] that encoded temporal information by tracking the SIFT. This effectively compressed the size of the SIFT features. The work by [7] used the fast CenSurE keypoint detector and BRIEF descriptor instead of SIFT. Binary Temporal Alignment was used to efficiently find a match. The MPEG-CDVS Standard [8] adopted the scalable compressed Fisher Vector (SCFV) representation for visual search. The SCFV achieved high matching accuracy with minimum memory requirements. Inspired by [8], we adopt fisher vector aggregation in our work for generating our deduplication feature.

In this work, we propose a robust and distinctive deduplication feature for finding near-duplicates among a large video repository. The feature is computed in three steps. First, SIFT keypoints are extracted from uniformly-sampled video frames which are used for training Gaussian Mixture Models (GMM). Next, GMM is used for fisher vector aggregation. Lastly, the fisher vector computed is used to generate a deduplication feature by passing through a lightweight Multilayer Perceptron (MLP) network. The proposed feature is lightweight in terms of time consumption. The feature proposed was evaluated on the CC_WEB_VIDEO [9] dataset. The main contributions of this paper are:

- We propose a robust and discriminative deduplication feature for searching duplicates/near-duplicates for the task of video deduplication.
- The proposed deduplication feature is lightweight as compared to deep neural networks like CNN.

The remainder of the paper is organized as follows. The proposed method is described in Section 2. Experimental results are presented in Section 3. Conclusion is presented in the last section of this paper.

This work is partially supported by NSF grants 1747751 and 2148382.

2. PROPOSED METHOD

In this section, we describe the proposed approach for near-duplicate image retrieval. The overall workflow of the proposed system is illustrated in Fig. 1. The proposed method consists of the following steps:

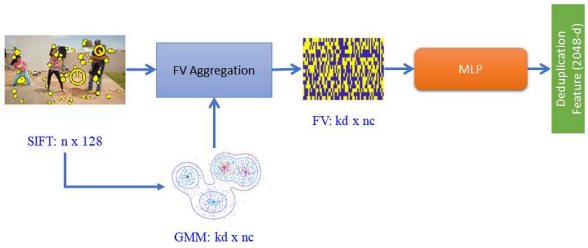


Fig. 1: Overall workflow of the proposed system.

2.1. Fisher Vector Aggregation

Fisher vectors assign local descriptor features (SIFT keypoints in our case) to elements in a probability visual vocabulary. The probability visual vocabulary is obtained via a generative model. We use the Gaussian Mixture Models (GMM) as the generative model for fisher vector aggregation. SIFT keypoints were used as features due to its scale and rotation invariance.

We fit the GMM using SIFT keypoint descriptors [3] which were extracted from frames in the CDVS dataset [8]. The dimension of each SIFT keypoint descriptor is 128-d. We reduce the 128-d SIFT keypoint descriptor into a 16-d and 24-d descriptor before using it for training a GMM. Principal Component Analysis (PCA) was used to project the 128-d descriptor to a 16-d and a 24-d descriptor. To train the GMM, we use the projected SIFT keypoints randomly sampled from the entire dataset. We train two GMM - one using 16-d descriptor and one using 24-d descriptor. The outcome of fitting the GMM is a visual vocabulary of dominant image features and their distributions.

Using the SIFT keypoint descriptors and the two fitted GMM, two fisher vectors are generated. The dimensions of the generated fisher vectors are 2048-d and 3072-d for 16-d and 24-d trained GMM, respectively.

2.2. Deduplication Transfer Learning with MLP Feature

The two fisher vectors generated via the procedure explained in the previous subsection are used as inputs to train a lightweight Multilayer Perceptron (MLP) network. The MLP produces the final robust and discriminative feature that will be used for finding duplicates/near-duplicates.

We propose to use a simple 7-layer MLP network that is lightweight as compared to a Convolutional Neural Network (CNN). The architecture of the MLP can be visualized in Fig. 2. The first two linear layers are followed by Parametric Rectified Linear Unit (PReLU). Sigmoid activation function is used before the last linear layer (FC-layer) of the MLP. After training the MLP, we remove the FC-layer and extract an embedding of dimension 2048-d. This is the feature which we use for video deduplication. Similar to [10, 11], we use a combination of center loss and softmax loss (see equations 1-3) for training the MLP.

$$L_{center} = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{1}$$

$$L_{softmax} = -\sum_{i=1}^{m} \log \frac{e^{W_{l_i}^T \hat{x}_i + b_{l_i}}}{\sum_{j=1}^{n} e^{W_j^T \hat{x}_i + b_j}} \tag{2}$$

$$L = L_{softmax} + \lambda \cdot L_{center} \tag{3}$$

Fig. 2: Network architecture for the proposed MLP.

2.3. Duplicate Verification

In this step, we propose a simple yet effective algorithm to search for duplicates/near-duplicates using the proposed deduplication feature. A pictorial explanation of the algorithm can be seen in Fig. 3. Given 2 video sequences (Seq_q and Seq_s), of equal or unequal length, we compute the similarity score (S_{final}) between the sequences as a function of average of off-setted matching distance. Sequences are represented as frames sampled at uniform intervals. It is assumed that the sequences are uniformly sampled at fixed intervals. We will have two cases - one in which the two sequences are of equal length and another in which the lengths are unequal. Cosine similarity measure is used for computing the similarity score. It is defined as:

$$S_c(A,B) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

where A_i and B_i are components of vectors A and B, respectively.

For the former case, we compute frame-to-frame cosine similarity and average the similarity scores to obtain a single similarity value S_{final}. For example, if Seq_q and Seq_s contain n frames each, we compare the n-th frame in Seq_q with the n-th frame in Seq_s, $n+1$-th frame from Seq_q to $n+1$-th

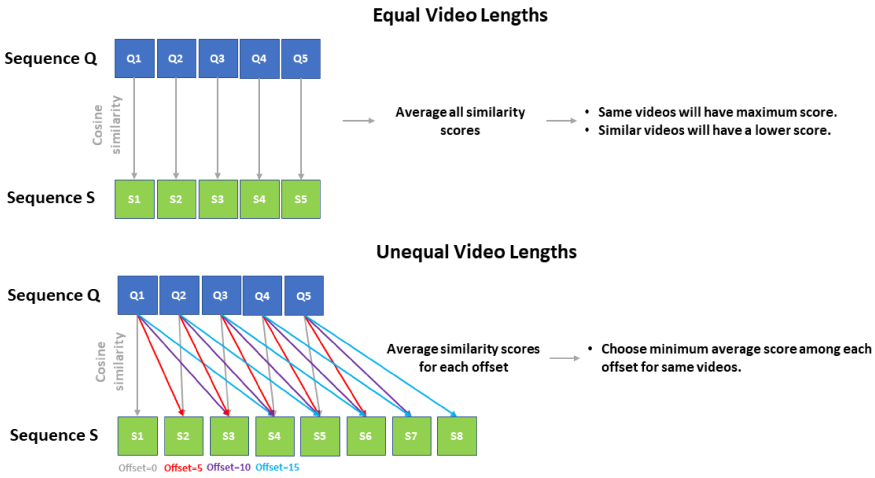


Fig. 3: Algorithm for the duplicate verification step. Offset is equal to the step size while uniformly sampling the frames from videos.

in Seq_s and so on. S_{final} is obtained by averaging all the similarity scores.

In case of unequal lengths, we need to calculate the offset and choose the minimum of the average of similarity scores as S_{final}. Let the length of Seq_q be $l_q = 5$ and that of Seq_s be $l_s = 8$. We take the smaller sequence (Seq_q) and compare its n-th frame to the n-th frame in Seq_s upto l_q. This process is repeated by starting the comparison again starting from $n+1$-th frame in Seq_s. Each iteration creates an offset value which increases based on the granularity of the frame sampling. The offset with the minimum average of similarity scores will be chosen as the S_{final}.

3. EXPERIMENTS

3.1. Dataset Details

The Vimeo90k [12] and CC_WEB_VIDEO [9] datasets were used for training and testing of our video deduplication system, respectively.

- Vimeo90k dataset: It contains 89,800 videos downloaded from vimeo.com. We use the 'Triplet dataset' available with the Vimeo90k dataset for the task temporal frame interpolation. The triplet dataset contains 73,171 3-frame sequences. Sequences are extracted from 15k selected videos from the Vimeo90k dataset and each sequence has a resolution of 448 x 256. Vimeo90k dataset was used for training the MLP.

- CC_WEB_VIDEO dataset: This dataset consists of 24 queries from YouTube, Google Video, and Yahoo Video. It contains 3401 near-duplicate videos which are almost identical to the exact duplicate of each other. The variations occur in file formats, encoding parameters, photometric variations, editing operations, etc.

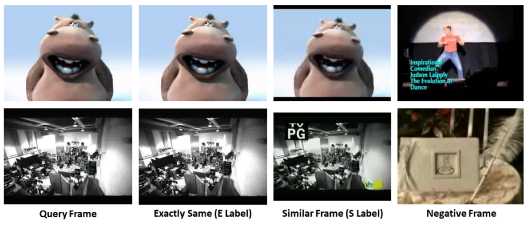


Fig. 4: Samples from the CC_WEB_VIDEO [9] dataset.

These 3,401 videos only contain videos with labels 'E' and 'S' which represent 'exactly duplicate' and 'similar video', respectively. This dataset was used for testing our deduplication feature.

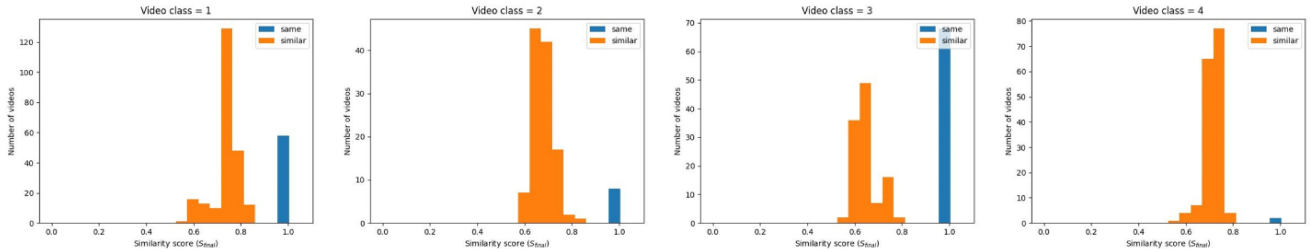
3.2. Implementation Details

The system was implemented using MATLAB and python programming languages. The extraction of SIFT keypoints features, GMM training, and computation of fisher vector were implemented using MATLAB. The MLP and testing of the deduplication feature were implemented using Python. PyTorch framework was used for implementing and training/testing the MLP.

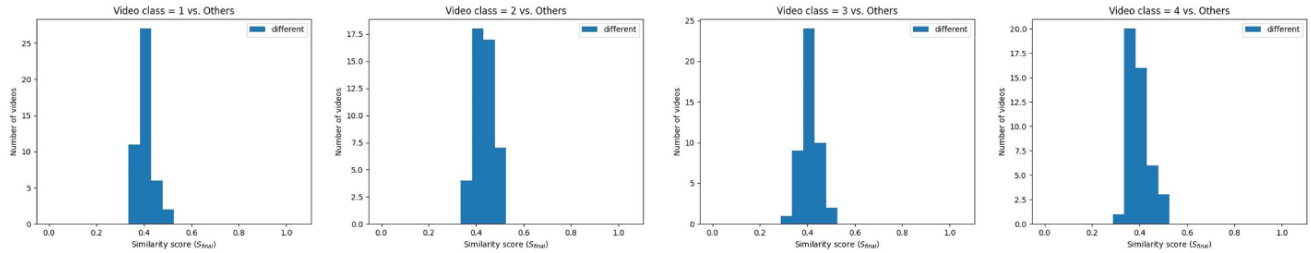
3.3. Experimental Results

This subsection validates the robustness and effectiveness of our proposed deduplication feature for video deduplication. We present results for three experiments which are explained in the following text. Due to the limited number of pages, for the first 2 experiments, we only show the results for the first 4 classes in the CC_WEB_VIDEO dataset.

Intraclass feature verification: In this experiment, we test the proposed feature by computing the S_{final} score for



(a) Histogram plot for intraclass experiment.



(b) Histogram plot for interclass experiment.

Fig. 5: Histograms for intra and inter class feature verification.

features of exactly duplicate (E label) (see Fig. 4) and similar videos (S label) (see Fig. 4) in the CC_WEB_VIDEO dataset for the same class. Ideally, the S_{final} score between a query video and its exact duplicate videos should be 1 whereas the distance between the query video and its similar videos should be < 1. Each video class has a seed video and we use this seed video from each class to compare it with all other videos in the same class. The histogram of S_{final} similarity scores for this experiment can be seen in Fig. 5a. As evident in Fig. 5a, the seed-to-exact duplicate score is 1 and the seed-to-similar score is near < 1.

Interclass feature verification: The second experiment was conducted to test the similarity score (S_{final}) between seed video from each class to all other videos in the other classes. Ideally, this score should be much lower than the seed-to-similar score (from preceding section). Particularly, we generate 46 pairs of frames randomly from each query class. Each pair includes frame from the query class and another frame from other classes (negative frame see Fig. 4). None of the pairs contain frames from the same class. The results for this experiments can be seen in Fig. 5b. It can be seen that the score is indeed lower than that of the seed-to-similar score.

Intra/Inter-class feature verification: In this experiment, we test the effectiveness of the proposed feature for both intra and inter class videos. Fig. 6 shows the Receiver Operating Characteristic (ROC) curve. For the blue curve in Fig. 6, we take seed video from each class and compare it to all E and S labeled frames within the same class. Also, the seed video is compared to negative videos. In case of the orange curve in Fig. 6, seed video from each class is compared to S labeled videos from the same class and to negative videos. These results validate the robustness of our proposed deduplication feature.

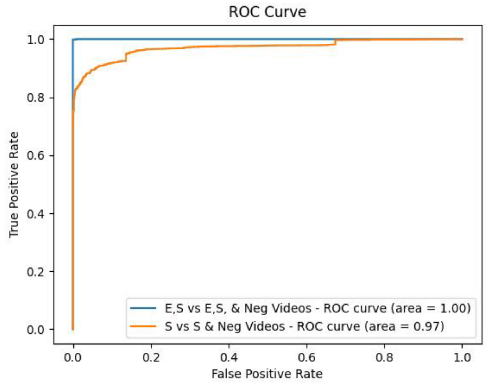


Fig. 6: ROC Curve for intra/inter-class feature verification.

4. CONCLUSION

In this paper, we present a lightweight and robust deduplication feature for finding duplicate/near-duplicate videos. The solution proposed is based on fisher vector aggregation and a lightweight multilayer perceptron (MLP). The fisher vector aggregation uses Gaussian Mixture Models (GMM) as the generative model. The GMM was trained via SIFT keypoints as input. The robustness of the feature was tested on the CC_WEB_VIDEO dataset. The results confirm that the proposed feature is invariant to typical editing effects.

5. REFERENCES

[1] Matthijs Douze, Hervé Jégou, and Cordelia Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Transactions on Multimedia*, vol. 12, no. 4, pp. 257–266, 2010. 1

[2] Krystian Mikolajczyk and Cordelia Schmid, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004. 1

[3] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. 1, 2

[4] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid, "Description of interest regions with local binary patterns," *Pattern recognition*, vol. 42, no. 3, pp. 425–436, 2009. 1

[5] Yu-Gang Jiang, Yudong Jiang, and Jiajun Wang, "Vcdb: a large-scale database for partial copy detection in videos," in *European conference on computer vision*. Springer, 2014, pp. 357–371. 1

[6] Yingying Zhu, Xiaoyan Huang, Qiang Huang, and Qi Tian, "Large-scale video copy retrieval with temporal-concentration sift," *Neurocomputing*, vol. 187, pp. 83–91, 2016. 1

[7] Yue Zhang and Xinxiang Zhang, "Effective real-scenario video copy detection," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 3951–3956. 1

[8] Ling-Yu Duan, Vijay Chandrasekhar, Jie Chen, Jie Lin, Zhe Wang, Tiejun Huang, Bernd Girod, and Wen Gao, "Overview of the mpeg-cdvs standard," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 179–194, 2015. 1, 2

[9] AGH Xiao Wu, Chong-Wah Ngo, and A Hauptmann, "Cc web video: Near-duplicate web video dataset," . 1, 3

[10] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, pp. 107069, 2020. 2

[11] Rijun Liao, Weizhi An, Zhu Li, and Shuvra S Bhattacharyya, "A novel view synthesis approach based on view space covering for gait recognition," *Neurocomputing*, vol. 453, pp. 13–25, 2021. 2

[12] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019. 3